IS IT REALLY FUN? DETECTING LOW ENGAGEMENT EVENTS IN VIDEO GAMES

10/23/25

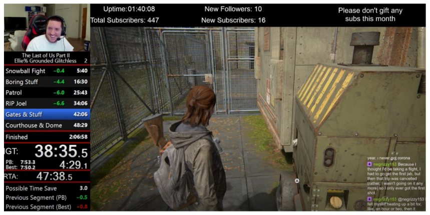
Zhiao Wei

Problem exsisting

video games requirement: They need to entertain the users and be fun to play.

Laugh ⇏ engagement

Key problem: video games are interactive



year. i never got corona

negrizzy153 Because I thought I'd be taking a flight, I had to go get the first jab, but then that trip was cancelled (rather, I wasn't going on it any more) so I only ever got the first shot.

negrizzy153 @negrizzy153 felt myself heating up a bit for, like, an hour or two, then it

Fig. 1: Example of low-engagement

How they solved?

Datapoints（40 players × 8 games × 3 minutes per game × 1 annotation per minute = 1,130 data points）

↓

Method 1: Affectiva（Business sentiment analysis tool based on 10 million user data. Testing 10 thresholds to find the optimal one.）

Method 2: K+ （ If neutral sentiment dominates → low engagement ）

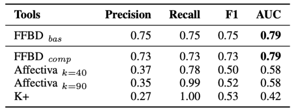
Method 3: FFBD（Facial Features-Based Detection）

↓

Results

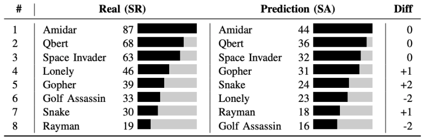
Results - 1

TABLE IV: Comparison between FFBD and the other two approaches.



Tools	Precision	Recall	F1	AUC
FFBD $_{bas}$	0.75	0.75	0.75	**0.79**
FFBD $_{comp}$	0.73	0.73	0.73	**0.79**
Affectiva $_{k=40}$	0.37	0.78	0.50	0.58
Affectiva $_{k=90}$	0.35	0.99	0.52	0.58
K+	0.27	1.00	0.53	0.42

TABLE VI: RQ$_2$: Video Games rankings (direct and predicted), with the number of low engagement events for both the scenarios and the rank difference between the two.

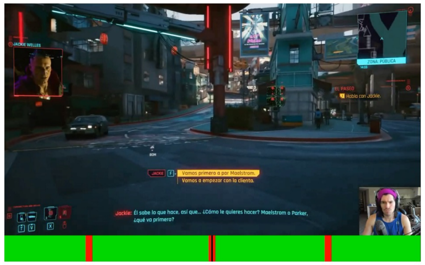


#	Real (SR)		Prediction (SA)		Diff
1	Amidar	87	Amidar	44	0
2	Qbert	68	Qbert	36	0
3	Space Invader	63	Space Invader	32	0
4	Lonely	46	Gopher	31	+1
5	Gopher	39	Snake	24	+2
6	Golf Assassin	33	Lonely	23	-2
7	Snake	30	Rayman	18	+1
8	Rayman	19	Golf Assassin	16	-2

- FFBD achieved 74.7% accuracy.

- FFBD can accurately identify the top 3 most boring games, and the overall ranking is highly consistent with players' subjective evaluations.

Results – 2 industry applicability



Vamos primero a por Maelstrom.
Vamos a empezar con la clienta.

Jackie: Él sabe lo que hace, así que... ¿Cómo le quieres hacer? Maelstrom o Parker, ¿qué va primero?

Fig. 2: An example of what we showed to participants. The bar below shows engagement in time (red → low engagement).

- It detected at least 5 low engagement events.
- **red**: potential low engagement events
  green: the non-low engagement events

Positives

- Practical Application for game designer.
- Post-Release Monitoring.
- Fine-Grained Real-Time Detection.

Negatives

- Low Recall Rate. The model only achieves 41% recall for the low engagement class.
- Cannot Distinguish External Factors. Like chat interaction, donations.
- Limited Information Dimensions. Maybe add streamer's audio commentary, in-game actions, and chat interactions.

Possible Future Work / Rating

1. Multimodal Data Integration. Incorporate additional data modalities to improve detection accuracy.
2. Expanding Dataset Scale. Collect data from a more diverse population.

Rating:(3/5): This is a high-quality empirical paper that pioneers real engagement detection with promising results (83.3% ranking correlation) and positive industry feedback, but is held back from top-tier status by low recall (41%), limited dataset scale (40 players, 8 games), and lack of causal insights into *why* engagement drops.

Discussion Points

1. Is low engagement always a design flaw, or could it be player factors, external distractions, or intentional design (e.g., tension in horror games)?
2. Does "real engagement" measured with constant interruptions actually reflect uninterrupted gameplay engagement? Can a model trained on short, fragmented lab sessions generalize to long natural streaming sessions?
3. Should we prioritize recall (catching all issues, currently only 41%) or precision (reducing false alarms, currently 74.7%) for practical use? In industrial settings, which metric is more critical?

UNIVERSITY OF WATERLOO

FACULTY OF MATHEMATICS

Thank you!